

Modeling the demand for train kilometers: A microeconomic approach

– draft –

Daniel J. van Vuuren Piet Rietveld

Vrije Universiteit, Amsterdam and Tinbergen Institute
Faculty of Economics, Business Administration and Econometrics
De Boelelaan 1105
1081 HV Amsterdam
e-mail: dvuuren@tinbergen.nl

January 31, 2000

Abstract

Consumer demand for rail transportation has traditionally been analyzed by means of aggregate demand systems and disaggregate discrete choice models. It is remarkable however that no serious efforts have been made to develop a disaggregate structural demand model, which takes account of the fact that consumers face a nonlinear budget constraint. It is argued that the use of such a model is necessary, because individuals typically have the opportunity to choose between many different types of tickets. It is therefore clear that consumer demand for transportation not only depends on price, but also that the ‘consumption’ of a certain amount of transportation will have causal influence on price. An important distinction between the present case and earlier studies of ‘discrete/continuous goods’, such as labor supply and electricity demand, concerns the nature of the discrete choice: While in earlier applications one single simultaneous choice is made for both the discrete and the continuous choice, the demand for transportation requires two explicit choices - a discrete choice for mode and/or ticket type and a continuous choice for the amount of transportation. Evidence from our data suggests that the explicit nature of the discrete choice is likely to lead to an extra source of optimization error as compared to Hausman’s [9] overview, which states that many observed combinations of discrete and continuous choice are demonstrably suboptimal - regardless of individual preferences. Estimation of the model for non-peak hour travelers by train in the Netherlands shows that the (absolute value of the) price elasticity of the demand for train kilometers equals about unity, and that the income elasticity is fairly small (about 0.05). It is suggested that the estimated model can be extended in many ways.

1 Introduction

In microeconomic textbooks (e.g. Varian [22]) it is mostly assumed that goods can be consumed in continuous quantities. For the case of travel demand this does not look like an unrealistic assumption, as the number of kilometers or the number of trips can be reasonably considered as continuous variables. Another common assumption is that price is constant, regardless of the quantity that is consumed. This latter assumption is not realistic however in the case of travel demand - either per rail or for other modes of public transport. Individuals are able to make choices between different types of train tickets and are therefore, to a certain extent, able to determine their own unit price. It is this discrete choice that makes the analysis deviate from the textbook case. It is remarkable however that there has been only scant attention for discrete/continuous analyses of the demand for transport services, despite the awareness that such analyses are most suitable for this problem (Hensher and Milthorpe [11]). An exception is De Jong's [5] analysis, who examines the simultaneous household demand for automobiles and automobile kilometers.

Traditionally, the analysis of consumer demand of discrete/continuous goods has focused on the case of labor supply (amongst others Burtless and Hausman [3] and Arrufat and Zabalza [1]). In these studies the continuous good is hours of work (or hours of leisure) and the discrete choice concerns the wage rate. It is well-known that in most countries tax rates are not constant, but are in general varying (mostly increasing) with the wage rate. This is the cause of consumers facing a nonlinear budget constraint, which implies that the analysis has to take the discrete choice explicitly into account, as neglecting this would yield biased estimates. (Comparisons of different estimators can be found in Herriges and King [12], Blomquist [2] and Ericson and Flood [7]). Apart from the literature on labor supply there are many other empirical studies that have been focusing on the consumer demand for goods that have a discrete as well as a continuous component. Examples can be found in the area of *electricity demand* (Dubin and McFadden [6] and Herriges and King [12]), *water demand* (Hewitt and Hanemann [13]), and the *demand for housing* (King [15], Hausman and Wise [10], Hoyt and Rosenthal [14]). Important review papers in this area of the demand of 'discrete/continuous goods' have been written by Hanemann [8], Hausman [9] and Moffitt [19].

An important distinction of the current work with most earlier studies concerns the nature of the discrete choice. In many applications, like e.g. labor supply, the discrete choice is automatically tied to the continuous choice and therefore has an implicit nature. For example, a person who chooses to work H hours a week, will automatically face the tax rate that goes together with this amount H . In other applications however, such as the demand for telecommunication, the demand for housing and travel demand, the discrete choice has a more explicit

nature. Whereas from an economic point of view this issue seems not relevant (due to the assumption that consumers behave as ‘rational’ utility maximizers), it can be argued that from an econometric point of view the distinction between these two natures of discrete choice is very important. In particular, the inclusion of optimization error into the model requires special care, as its nature is closely linked to that of the discrete choice. One may expect, for example, that the magnitude of the optimization error is larger in the case of an explicit choice as there exists a possibility that people make the ‘wrong’ discrete choice in combination with their continuous choice. For example, a person who has decided to buy a reduced fare pass and to travel at reduced fare, may travel less kilometers than he had foreseen. Therefore it may turn out that he has made the wrong choice of travel card. A study that is closely linked to this issue has been conducted by Train et al. [20], who analyze the relation between tariff choice and consumption of transport kilometers from a purely discrete choice framework. Indeed the authors have found that this optimization error plays a more important role if consumers have to make an explicit choice about the discrete part of their demand.

In section 2 we discuss a demand model for the combined demand for train kilometers and train tickets. The empirical part of the paper consists of section 4. It estimates the proposed model with the data that is briefly discussed in section 3. Section 5 concludes.

2 A structural model for the demand for train kilometers

Following the most elementary of microeconomic theory, the individual’s demand can be written as a function of income and price. A possible specification of such a demand model would be the log-log specification:

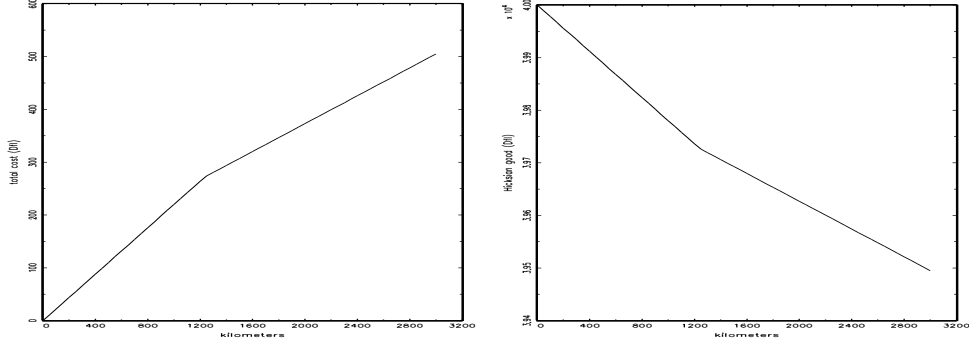
$$\ln k = \alpha \ln y + \beta \ln p + \delta, \quad (1)$$

where train kilometers are denoted by k , income is denoted by y and the price of a train kilometer is denoted by p . Note that in this specification the income elasticity equals $\epsilon_y = \frac{\partial \ln k}{\partial \ln y} = \alpha$, and the price elasticity equals $\epsilon_p = \frac{\partial \ln k}{\partial \ln p} = \beta$, which explains why this model is often called the *constant elasticity model*. However, as has been known for a very long while, straightforward (OLS) estimation of this model would yield improper estimates, because the demand for kilometers has a causal influence on the kilometer price. Therefore, it is necessary to add an equation to the model that adequately describes the causality of kilometers on kilometer price. There are several ways to do this, but from the economist’s point of view it is certainly most appealing to let this additional equation be consistent with microeconomic theory¹.

The proper way of doing so starts by deriving the indirect utility function. This function can be derived by rewriting Roy’s identity as a differential equation and making use of the implicit

¹Another - more mechanical - approach has been proposed and estimated by van Vuuren and Rietveld [21].

Figure 1: Nonlinear price: Cost function and budget set



function theorem. A detailed account of this derivation is found in Burtless and Hausman [3]. For the constant elasticity specification in (1) the indirect utility function equals:

$$v(p, y) = \frac{1}{1 - \alpha} y^{1 - \alpha} - \frac{1}{1 + \beta} p^{1 + \beta} e^{\delta}. \quad (2)$$

Assume now that a consumer is able to choose between two types of train tickets: Full tariff tickets and reduced tariff tickets. If he chooses to travel at full tariff then he has to pay a price that is equal to p_f per kilometer. If, on the other hand the individual decides to buy a reduced fare pass at a fixed cost of p_r , then he may buy his travel tickets at a price of hp_f per kilometer, in other words he gets a $100(1 - h)\%$ reduction on his marginal price². As is depicted in Figure 1 this leads to a nonconvex budget constraint: Up to a certain ‘consumption’ one will choose to travel at full tariff, while after exceeding this level it becomes more profitable to travel at reduced tariff. The graphs are on a yearly basis, implying an average threshold level of about 1240 kilometers per annum. The first graph depicts the cost function as a function of kilometers while the second graph depicts the budget constraint for a net income of Dfl 40,000.

Whithin the microeconomic framework one will favor to travel at reduced fare if

$$v(hp_f, y - p_r) > v(p_f, y),$$

which in the current specification rewrites as

$$\delta > G(y), \quad (3)$$

²In the Netherlands one obtains a 40% reduction on his train ticket with a reduction pass, and hence $h = 0.60$.

where

$$G(y) := (1 - \alpha) \ln y - (1 + \beta) \ln p_f - \ln(1 - \alpha) + \ln(1 + \beta) + \ln \left[1 - \left(1 - \frac{p_r}{y} \right)^{1-\alpha} \right] - \ln \left[1 - h^{1+\beta} \right].$$

In the context of discrete/continuous choice models Hausman [9] distinguishes between four different sources of randomness:

1. *measurement error*. Possible sources are (i) the translation from yearly data to monthly data, (ii) the reported number of kilometers may be subject to bias due to unreported trips, (iii) the ‘measurement’ of income (see also section 4 and Appendix A).
2. *preference heterogeneity*. Of course it is not possible to include all determinants of what is usually referred to as ‘taste’. It is therefore appropriate to include a stochastic term that takes account of this unobserved variation in taste.
3. *optimization error*. For individuals, it may not always be possible to obtain the desired optimal package as a result of practical circumstances. (For example, people may get ill after having bought a reduced fare pass.) Secondly, people might make ‘errors’ in choosing an efficient combination of train kilometers and train tickets. This last phenomenon is discussed in more detail in section 3.
4. *specification error*. This is due to the modeler’s inability to specify a perfect econometric model.

One can imagine that with all these sources of stochasticity it is possible to construct a wide variety of econometric specifications that go together with (1). We propose to start with a simple two-error model:

$$\ln k = \alpha \ln y + \beta \ln p + \delta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \quad (4)$$

$$v(p, y) = \frac{1}{1 - \alpha} y^{1-\alpha} - \frac{1}{1 + \beta} p^{1+\beta} e^{\delta + \nu}, \quad \nu \sim N(0, \tau^2). \quad (5)$$

In this specification ε will contain both specification error and measurement error, while (stochastic) preference heterogeneity and optimization error would be a part of *both* ε and ν . Some might regard this as a disadvantage of the current specification, because they prefer to identify e.g. preference heterogeneity separately instead of the two combined error terms. On the other hand it may be possible to include more error terms into the model, but this may result in quite some additional computational work - in particular numerical integration routines.

The likelihood of owning a reduced fare ticket and traveling k kilometers equals

$$\ell_r := f \{ \varepsilon = \ln k - \alpha \ln(y - p_r) - \beta \ln(h p_f) - \delta; \nu > G(y) - \delta \}.$$

Under the preassumed distributional specifications for ε and ν this likelihood rewrites as

$$\ell_{r,k} = p_{k|r} \cdot P_r,$$

where

$$p_{k|r} = \frac{1}{\sigma} \phi \left(\frac{\ln k - \alpha \ln(y - p_r) - \beta \ln(h p_f) - \delta}{\sigma} \right),$$

$$P_r = 1 - \Phi \left(\frac{G(y) - \delta}{\tau} \right).$$

Similarly, the likelihood of traveling k kilometers at full tariff equals

$$\ell_{f,k} = p_{k|f} \cdot (1 - P_r),$$

where

$$p_{k|f} = \frac{1}{\sigma} \phi \left(\frac{\ln k - \alpha \ln y - \beta \ln p_f - \delta}{\sigma} \right).$$

Hence the log-likelihood for the whole dataset becomes

$$\ln \ell = \sum_{i: \text{"full tariff"}} \ln \ell_{f,k} + \sum_{i: \text{"reduced tariff"}} \ln \ell_{r,k}$$

As an extension of the current model one can think of

1. introducing correlation between ε and ν ;
2. decompose ε into a ‘structural’ and a ‘nonstructural’ part: $\varepsilon = \nu + \mu$;
3. introducing a random coefficients specification, cf. Bortolero and Hausman [3];
4. introducing an additional error term.

The latter improvement causes over-identification if an additive normal error term is included. It is however not unrealistic to include a nonnormal error term into the model, as the discrete choice optimization bias seems highly unsymmetric (see section 3). The second model can be considered as the ‘classical’ model as it has been estimated in practically all studies up to present. In this specification ν is defined as preference heterogeneity, while μ captures the other error terms. This is a logical motivation, as this preference heterogeneity is the only sort of disturbance that is known to the individual and thus is allowed to enter both his indirect utility function *and* demand function.

3 Data

In this paper we make use of a dataset from the Netherlands Railway (NS), the so-called 'Basisonderzoek' (BO). This dataset was obtained through a survey during the period April 1992 until March 1993. The most important objective of the study was to analyze the profiles of travelers, in particular those of travelers by train. The survey consisted of two parts: At first a random sample from the population of Dutch households was drawn, after which one individual from the household was asked about personal characteristics like age, gender and education. Moreover the subjects were asked about their possession of train fare reduction tickets, distance to the nearest train station and other factors of interest for the Dutch Railway. The second part of the survey concentrated on individuals who actually had been making use of the train as a means of transport. In this part every subject was asked to keep a month's diary of his travel behavior as far as railway transport was concerned.

In fact the dataset consists of two parts: The first part is a cross section dataset of the Dutch population. The subjects in the second dataset are a subsample of those in the first, satisfying two selection criteria: (i) the subject indicated that he traveled by train during the past year, and (ii) the subject was willing to keep a travel diary. For the purpose of this paper we have linked both datasets in order to be able to include individual characteristics into the demand equation that was specified in (1).

In Table 1 frequencies for "efficient" and "inefficient" travelers are reported, where an "efficient traveler" is defined as an individual who buys the right type of train ticket, given his demand for train kilometers. Formally, a traveler who owns a reduction ticket and has reported to travel k kilometers is called efficient if

$$c_f(k) \geq c_r(k),$$

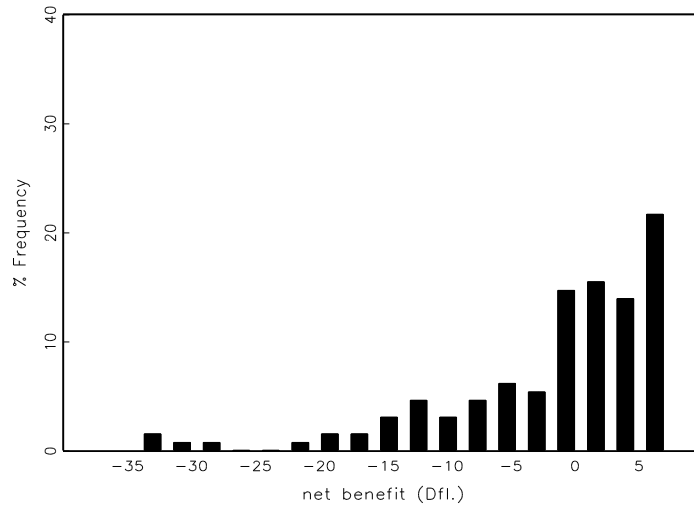
where c_f is the full tariff cost function and c_r is the reduced tariff cost function³. A similar definition can be given for "efficient reduced fare travelers". Figures 2 and 3 depict an absolute measure of efficiency, which is defined as the difference between the cost of the ticket type that is not chosen and the cost of the chosen ticket type. (For reduction card holders this benefit equals $c_f(k) - c_r(k)$.) On average, one is indifferent between the two types of tickets at a yearly 'consumption' of 1240 kilometers. The many negative values in Figure 2, some with considerable magnitude, as compared to the few negative values in Figure 3, all with low magnitude, suggest that many individuals are reluctant to buy a reduced fare ticket. It should be noted that the current findings bear much resemblance to Moffitt's [18] ascertained disutility from participation in a welfare program, and Koning's [16] finding that there is an 'entry fee effect' for households that might apply for rent assistance. For the current case, several explanations are possible for

³Note that in the model of section 2 this inequality is equivalent with (3).

Table 1: Budgetary efficiency of travelers (%)

| | efficient | inefficient | total |
|----------------|-----------|-------------|-------|
| full tariff | 29.6 | 28.3 | 58.9 |
| reduced tariff | 35.9 | 6.3 | 41.1 |
| total | 65.5 | 34.5 | 100.0 |

Figure 2: Full tariff travelers and optimality

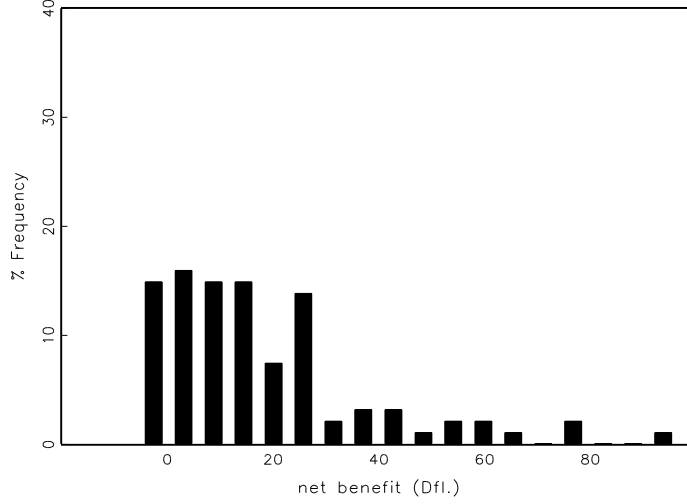


this ‘bias’, e.g. lack of information and application costs (mainly time). In addition one should note that the consumer’s decision to buy a reduced fare pass or not has to be taken *in advance*, which may cause a difficulty for certain individuals, who therefore ‘wait and see’ instead of taking preventive action (i.e. buying a reduced fare pass).

4 Estimation

In our estimations we have focused on persons that actually did have to make a choice between different types of tickets. This means that we have excluded groups which get their travel cards for free. In the Netherlands this mainly concerns students, people from the army and railway personnel (around 12% of the sample). Only Second Class passengers (about 90% of the sample)

Figure 3: Reduced tariff travelers and optimality



have been selected in the final dataset.

To allow for nonstochastic variation in individual taste the following specification for δ has been used:

$$\delta = \delta_0 + \delta_1 \text{AGE} + \delta_2 \text{CAR1} + \delta_3 \text{CAR2} + \delta_4 \text{HHSIZE} + \delta_5 \text{HWIFE}.$$

In this specification “CAR1” and “CAR2” are dummy variables which signify “can always make use of a car” and “can often make use of a car” respectively. Other included variables are “age”, “household size” and a dummy that indicates whether the individual is a housewife or not.

The actual fixed cost of a reduced fare card (p_r) equalled Dfl. 9 in 1992 and Dfl. 8.25 in 1993 (on a monthly basis). In order to avoid a new endogeneity problem, the “full tariff” kilometer price was fixed at the average (full tariff) kilometer price (Dfl. 0.22).

As no data has been available to us on the respondents’ income, we have instead made use of ‘predicted income’ which is calculated according to the specification in Appendix A. The bias that is due to this approximation flows into the error term ε in (4) as additional measurement and specification error.

The actual estimates can be found in Table 2. All parameter estimates are significant at a 5% confidence level; all the estimates have the sign that is consistent with prior expectations

Table 2: Model estimates, n=4373

| log-likelihood | -9433 | | -9227 | |
|------------------------|-----------|----------|-----------|----------|
| σ | 1.0574 | | 1.0469 | |
| τ | 0.3797 | | 0.4112 | |
| Parameters | Estimates | Std.err. | Estimates | Std.err. |
| constant | 2.1216 | 0.1786 | 2.2999 | 0.1966 |
| age | - | - | 0.0056 | 0.0006 |
| car available (always) | - | - | -0.1494 | 0.0200 |
| car available (often) | - | - | -0.0742 | 0.0243 |
| household size | - | - | -0.0347 | 0.0066 |
| housewife | - | - | -0.0379 | 0.0195 |
| ln income | 0.0477 | 0.0192 | 0.0498 | 0.0230 |
| ln price | -1.1765 | 0.0603 | -1.0323 | 0.0612 |

and (microeconomic) theory. The Slutsky condition

$$\beta < -\frac{\alpha pk}{y},$$

demanding that the own substitution effect of train kilometers is negative, was satisfied for *all* data entries⁴. The results indicate that the demand for train kilometers is fairly inelastic with respect to income, which is a bit remarkable for the demand during non-peak hours. One can expect a relatively large share of recreation trips as compared to the frequency of recreation trips during peak hours, which suggests a relatively high income elasticity. On the other hand people with high income might have little time to travel during non-peak hours, which may have a downward effect on the income elasticity. With values of -1.17 and -1.03 the price elasticity seems to deviate only little from unity (in absolute value).

5 Discussion and conclusion

In this paper we have proposed and estimated a model for the demand for train kilometers. The kind of approach that was used has been utilized previously in other areas of interest, such as labor supply and electricity demand, but has not received much attention in the area of travel demand. The current estimates show that the price elasticity equals unity (in absolute value) and that the income elasticity is fairly small at 0.05.

⁴Note that the additional restrictions as set out in MaCurdy et al. [17] do not apply here, as the nonconvex budget constraint implies that no ‘kink point solutions’ are attainable.

To conclude this paper we mention several improvements of the model that can be used in future analyses:

- Integration into a more complete model which involves the demand for train kilometers during non-peak hours. An interesting feature of such, more integrated, models would be the determination of substitution effects with other types of demands. Substitution effects may exist between First Class/Second Class demand, the mentioned peak hour/non-peak hour demand and possibly between various means of transport.
- Other demand specifications need to be estimated and compared with the current model.
- More sophisticated error structures need to be estimated and compared with the current findings. Some suggestions were mentioned in section 2.
- Heckman correction for selection bias. It is however not sure if selection bias exists in this model, as there may be opportunity for a generalization of Hanemann's [8] statement concerning multivariate normal and GEV distributed error terms⁵.
- Income might be estimated by a better model. (It is known from earlier work in this area that the education variable contains quite some measurement error. Even more seriously, this variable may be the cause of some "endogeneity bias" in the current specification.)

Acknowledgment

The authors are indebted to Aico van Vuuren (Vrije Universiteit, Amsterdam and Tinbergen Institute) for useful comments.

Appendix A. The income equation

In equation (6) we provide an OLS estimation for income. In this model the logarithm of income is specified as a linear function of education ('EDU'), age ('AGE') and gender ('FEM'). This last variable equals 1 if the individual is a female, and 0 otherwise. Standard errors are denoted in parentheses below the corresponding estimates. All parameters turn out to be highly significant at the 5% and 1% level.

$$\ln \text{INCOME} = 5.72 + 0.542 \ln \text{EDU} + 0.231 \ln \text{AGE} - 0.633I\{\text{FEM}\} \quad (6)$$

(0.172) (0.0401) (0.0399) (0.0256)

$$R^2 = 0.325$$

⁵A necessary and sufficient condition for avoiding sample selection bias in discrete/continuous choice models due to neglect of zero-observations is found in Chiang and Lee [4]

It is seen that ‘gender’ explains a large share of the variation in individual income: on the average the model predicts that women earn 47% less than the average male person. Of course ‘discrimination’ is not the only explanation for this gap, as in the Netherlands labor participation of women is lower and many women that do work have part time jobs.

References

- [1] J.L. Arrufat and A. Zabalza, *Female labor supply with taxation, random preferences and optimization errors*, *Econometrica* **54** (1986), no. 1, 47–63.
- [2] N.S. Blomquist, *Estimation methods for male labor supply functions; How to take account of nonlinear taxes*, *Journal of Econometrics* **70** (1996), 383–405.
- [3] G. Burtless and J.A. Hausman, *The effect of taxation on labor supply: Evaluating the Gary negative income tax experiment*, *Journal of Political Economy* **86** (1978), no. 6, 1103–1130.
- [4] J. Chiang and L.F. Lee, *Discrete/continuous models of consumer demand with binding nonnegativity constraints*, *Journal of Econometrics* **54** (1992), 79–93.
- [5] G.C. de Jong, *Some joint models of car ownership and car use*, Ph.D. thesis, Universiteit van Amsterdam, 1989.
- [6] J.A. Dubin and D.L. McFadden, *An econometric analysis of residential electric appliance holdings and consumption*, *Econometrica* **52** (1984), no. 2, 345–362.
- [7] P. Ericson and L. Flood, *A Monte Carlo evaluation of labor supply models*, *Empirical Economics* **22** (1997), 431–460.
- [8] W.M. Hanemann, *Discrete/continuous models of consumer demand*, *Econometrica* **52** (1984), no. 3, 541–561.
- [9] J.A. Hausman, *The econometrics of nonlinear budget sets*, *Econometrica* **53** (1985), no. 6, 1255–1281.
- [10] J.A. Hausman and D.A. Wise, *Discontinuous budget constraints and estimation: The demand for housing*, *Review of Economic Studies* **47** (1980), 75–96.
- [11] D.A. Hensher and F.W. Milthorpe, *Selectivity correction in discrete-continuous choice analysis*, *Regional Science and Urban Economics* **17** (1987), 123–150.
- [12] J.A. Herriges and K.K. King, *Residential demand for electricity under inverted block rates: Evidence from a controlled experiment*, *Journal of Business and Economic Statistics* **12** (1994), no. 4, 419–430.

- [13] J.A. Hewitt and W.M. Hanemann, *A discrete/continuous choice approach to residential water demand under block rate pricing*, Land Economics **71** (1995), no. 2, 173–192.
- [14] W.H. Hoyt and S.S. Rosenthal, *Capital gains taxation and the demand for owner-occupied housing*, Review of Economics and Statistics **72** (1990), 45–53.
- [15] M.A. King, *An econometric model of tenure choice and demand for housing as a joint decision*, Journal of Public Economics **14** (1980), 137–159.
- [16] R.H. Koning, *Essays on applied microeconometrics*, Ph.D. thesis, Rijksuniversiteit Groningen, 1995.
- [17] T. MaCurdy, D. Green, and H. Paarsch, *Assessing empirical approaches for analyzing taxes and labor supply*, Journal of Human Resources **25** (1990), 415–490.
- [18] R. Moffitt, *An economic model of welfare stigma*, American Economic Review **73** (1983), 1023–1035.
- [19] ———, *The econometrics of piecewise linear budget constraints*, Journal of Business and Economic Statistics **4** (1986), no. 3, 317–328.
- [20] K.E. Train, M. Ben-Akiva, and T. Atherton, *Consumption patterns and self-selecting tariffs*, Review of Economics and Statistics **71** (1989), no. 1, 62–73.
- [21] D.J. van Vuuren and P. Rietveld, *The demand for tickets and travel cards of railway travelers*, paper presented at ERSA 39th European Congress, Dublin, 1999.
- [22] H.R. Varian, *Microeconomic analysis*, 3rd ed., Norton, 1992.